



Repository Registries: Characteristics, Issues and Futures

Nikkia Anderson
Gail Hodge

CENDI Secretariat

November 17, 2014

CENDI 2014-1

Repository Registries: Characteristics, Issues and Futures

Prepared and Published by
CENDI Secretariat
c/o Information International Associates, Inc.
Oak Ridge, TN

CENDI is an interagency working group of senior scientific and technical information (STI) managers representing the leading U.S. Federal research and development agencies. As of September 2014, CENDI has 13 member agencies: the Departments of Agriculture, Commerce, Defense, Education, Energy, Health and Human Services, Homeland Security, and Transportation; the Environmental Protection Agency, National Aeronautics and Space Administration, National Archives and Records Administration, National Science Foundation, and the Library of Congress. CENDI is a volunteer-powered membership organization that serves the federal information community - that is, all those who create, manage, aggregate, organize, and provide access to federally-funded data and publications resulting from the nation's \$150 billion annual investment in federal R&D. Member organizations represent a cross-section of federal data and publication providers, including libraries, data centers, aggregators, information technology developers, and content management providers. CENDI exists to promote the success of its members' missions and, for more than 25 years, has provided a forum in which to address common interests through education, advocacy and joint projects that leverage scarce resources and specialty capabilities.

Copyright Notice

This is a work of the U.S. Government and is not subject to copyright protection in the United States. Foreign copyrights may apply.

November 2014

Contents

Introduction	5
What is a Repository Registry?.....	6
Registry Examples	6
Registry Characteristics.....	7
Metadata	7
Ingest Practices and Workflows	7
Functionality	8
Issues and the Future of Registries	8
Recommendations for the CENDI Work Plan	9
 Appendix A: Repository Registry Descriptions by Type	 10
Registries of Open Access Publication Repositories	10
Registry of Data Repositories.....	13
Registry of Enterprise Repositories	15
Registries of Repositories Based on a Specific Technology	16
 Appendix B: Metadata by Registry	 19

Page intentionally left blank for two-sided copying.

Introduction

The Open Government/Open Data Initiative and call passed down by the Whitehouse Office of Science and Technology Policy (OSTP) states that “digitally formatted scientific data resulting from unclassified research supported wholly or in part by Federal funding should be stored and publicly accessible to search, retrieve, and analyze.”¹ Based on these directives, the government is requiring increased compliance with public access and transparency, producing the need for repositories where the results of government funded research can be deposited, discovered, and made accessible for re-use.

The US is not alone in its push for public access to research results. There is an international need to provide open access to scientific data and research. Mandates similar to that of OSTP are being passed in other countries; the European Commission recommends that member states create clear policies for open access that require “research data that result from publicly funded research become publicly accessible, usable and re-usable through digital e-infrastructures.”² Many of the open access registries that are currently available are run by the universities in the United Kingdom (UK), and there are similar activities in EU countries.

Although many federal government repositories meet the requirements from OSTP, it remains challenging for citizen, public, and private sector researchers to find and select the appropriate repository that meets the needs of the agency and the user. A repository registry provides a way for users to identify one or more repositories that could be used as a store house for agency data/research. This report discusses repository registries in general but also focuses on those currently available registries that point to repositories from federal government agencies.

Often authors, whether government employees, grantees or contractors, are unaware of what is available in their own agencies let alone in other agencies. Identifying an appropriate repository for deposit through a repository registry can:

- Help agencies identify repositories into which their authors can deposit data;
- Help name those repositories in the policies and procedures for authors;
- Identify a repository whose architecture could be leveraged to reduce the work it takes to build a repository from scratch.

While many registries might qualify as places for deposit of federally funded data sets, decisions on which do and do not meet agency requirements make a case for the study of criteria for trusted repositories. Criteria for evaluating repositories will be taken up in a related CENDI initiative, because ultimately both the identification and the vetting of repositories will be critical to data management planning (DMP).

¹ OSTP Public Access Memo 2013: Increasing Access to the Results of Federally Funded Scientific Research
http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf

² European Commission: Commission Recommendation of 17.7.2012 on access to and preservation of scientific information.
http://ec.europa.eu/research/science-society/document_library/pdf_06/recommendation-access-and-preservation-scientific-information_en.pdf

What is a Repository Registry?

A repository registry is a web based metadata catalog that points to repositories. The repositories are collections of data objects, such as documents, videos, audio files, datasets, etc., which may also contain metadata for each object. The registry allows users to search and browse metadata describing the repositories, which aids in identifying repositories that can be used to deposit data or from which data can be accessed. In general, registries are characterized by extensive metadata describing the repositories; a separation between the metadata about the repository and the repository content, which is usually hosted elsewhere; and a database or other structure for the metadata that allows for search functionality (including filtering) in addition to browse capabilities.

Registries are closely related to catalogs, directories and inventories, and these terms are often used interchangeably. However, unlike registries, they provide direct access to individual data objects, rather than to a series of repositories of data objects. The metadata describes the individual data objects. Some catalogs have extensive metadata, such as geospatial data catalogs. Directories and inventories have more limited metadata and are often simpler in structure. For example, a directory or inventory may be as simple as a list of links on an HTML page and lack a searchable database or other structured metadata.

Registry Examples

There are several major registries available. The list of registries below is not comprehensive, but selective to highlight registries that include U.S. Federal Government repositories. In most cases, the U.S. government repositories are a part of larger repository collections that include repositories from the private sector or other non-U.S. governments. The registries highlighted in this section are divided into those that focus on 1) open access publications, 2) data, 3) repositories from a particular enterprise or government, and 4) those based on a particular technology. The Repository Registries identified and described are:

- **Registries of Open Access Publication Repositories**
 - ROAR (Registry of Open Access Repositories) - <http://roar.eprints.org/>
 - OpenDOAR (Directory of Open Access Repositories) - <http://www.opendoar.org/index.html>
- **Registry of Data Repositories**
 - re3data.org (Registry of Research Data Repositories) - <http://www.re3data.org/>
- **Registry of Enterprise Repositories**
 - Digitization Projects Registry (GPO) - <http://registry.fdlp.gov> [Currently unavailable]
- **Registries of Repositories Based on a Specific Technology**
 - DuraSpace Registry - <http://www.duraspace.org/registry>
 - OAI Registered Data Providers - <http://www.openarchives.org/Register/BrowseSites>

Appendix A provides information about each registry including a list of the metadata, a description of the process of ingesting the metadata into the registry, and a list of federal agency repositories included in the registry. Each description is followed by a screen shot of the registry homepage. More information is available by following the link to the registry. See the registry URL to browse and/or search the collections.

Registry Characteristics

This section analyzes the registries highlighted in this report, looking at characteristics such as metadata, ingest processes and workflows, and user interface functionality.

Metadata

The metadata used to describe the repositories in a registry is a key factor in findability. There are commonalities and differences across the metadata elements used in the registries listed above.

Appendix B combines all the metadata elements from the registries in Appendix A, groups them by general category, and indicates the registries in which the element occurs. Note that the names may be different, but the type of information is the same. For this paper, we have not made that differentiation, nor have we attempted to reconcile the metadata elements. Instead, the metadata elements are grouped under general categories, such as Title, Identifier, etc. In addition, even similar elements may have different content. For example, the list for repository types is more extensive in the re3data.org registry versus the open access repositories (ROAR and OpenDOAR). This is likely due to the broader scope of re3data.org and the fact that it isn't limited to open access repositories only.

The categories created in Appendix B begin to identify some of the potential common elements used to describe repositories across registries, such as Repository Name, System/Registry ID, Location, Software Platform/Repository Technology, Repository Type, and Repository URL. A more thorough analysis of the content and definitions of the metadata elements in each registry would be needed to identify a core schema for the description of a repository.

Ingest Practices and Workflows

The need to keep the registries up to date is important. From the user perspective, it shows that the registry is a valuable resource. Most of the registries above accept repository suggestions via a web form of descriptive and administrative metadata fields. The metadata is used to evaluate the repository for possible inclusion. Some of the repositories are added to the registries through user account entry and automatic harvesting through a service such as OAI-PMH. ROAR metadata provides information on other registries that have the same repository cataloged. The identified registries are ingesting these records through metadata harvesters.

To make updates to some of the registries, there is a link provided on the record page for the repository owner to send or log-in to make updates. Duraspace provides an email link from each cataloged repository record requesting additions or updated information. OAI allows the maintainer of the repository to run the validation/registration process again to match new information that can be pulled automatically from the OAI Identify Response call. The OpenDOAR registry allows any user to suggest an update to a record by clicking a link from the browse/search results under the OpenDOAR ID.

Functionality

All registries have a browse-able interface. The common filter options include country, repository type, subject, and software. In addition to a browse function, most registries also have simple search capabilities. The re3data.org registry is unique in that it uses icons to identify when a registry is open access, has certificates, has been reviewed by re3data.org reviewers, and/or assigned a persistent identifier. These unique image icons make it easy for users to browse without using direct filters.

The displays of the results vary across the registries but a link to the repository is provided from each registry. The ROAR also links to other registries to point to the same repository.

Issues and the Future of Registries

The goal of a repository registry is to have a robust system with an extensive catalog within the scope of the registry's mission. Collection development is key to the sustainability of a registry. Continuing to improve and enhance the registry, shows why it is a valuable resource for depositors, funding organizations and users.

However, the potential proliferation of repositories based on open/public access requirements may lead to a similar proliferation of registries exacerbating the difficulties in finding the appropriate repository. Working groups within the Research Data Alliance (RDA), such as the Data Description Registry Interoperability Working Group, are addressing the problem of cross-platform discovery through a series of bi-lateral information exchange projects and software solutions. They are relying on other groups within the RDA to address the issue of Identity Awareness of Research Data which plays into cross-platform discovery.

The efforts underway by the RDA working groups and the current goals of the registries described in this report highlight the need for more visibility for the growing number of repositories. As the number of repositories within a registry increases, the organization of the web sites and the functionality of the databases and finding aids will need to keep pace. Browsing will require improved taxonomies to support user friendly organization of the sites and targeting of content. Search functionality will need to consider ways to refine, limit, filter and rank search results.

The need to harvest metadata for repositories raises the issue of standard metadata for describing repositories and consistent processes for alerting the registries when repository changes occur.

Registries may describe the policies used to “select” repositories for inclusion. However, they rely on descriptive information to support the evaluation of the repositories by those who are looking to deposit data or those looking to use them. As the number of repositories increases and the policies around public access to data of all types matures, registries will begin to use certification or evaluation criteria as part of the selection process and identify certifications or other indicators of quality and trustworthiness as part of the metadata.

Recommendations for the CENDI Work Plan

Given the requirements for Data Management Plans (DMP), there is increasing need to effectively identify registries for government use. It is recommended that CENDI should:

1. Investigate creating a Registry of SciTech and supporting repositories to help identify repositories that could be used for deposit in meeting DMP requirements.
2. Identify essential trust factors and other characteristics as the criteria for inclusion.
3. Continue to follow the development of standards and best practices as relates to repository registries particularly with regard to a core set of metadata for describing repositories.
4. Consider linkages to Science.gov.
5. Give serious consideration to having a Science.gov Registry (not just a catalog – see definitional differences) for publication repositories and possibly for data sets. This could be an enhancement to the current list of Science.gov authoritative databases (www.science.gov/searchdbs.html).

Appendix A: Repository Registry Descriptions by Type

Registries of Open Access Publication Repositories

ROAR (Registry of Open Access Repositories) http://roar.eprints.org/	
<i>A registry of open access repositories. Its goal is to promote open access by providing timely information about the growth of Open Access repositories.</i>	
Hosted/Maintained by:	University of Southampton, UK – School of Electronics and Computer Science
Funding by:	JISC
Part of :	Eprints.org
Level of Content:	Points to Repository Web site
Listed Metadata: *Filter options	ROAR ID Home Page Repository Type* Organization Software* Country* Location Subject* Birth Date Record Count Daily Deposit Activity OAI-PMH Interface Other Registries (ROARMAP; Celestial; OpenDOAR) Record Creator
Federal Repositories Included:	AgSpace (USDA NAL Digital Collections) Socrata (Native repository of data.gov) National Advisory Committee for Aeronautics (available from NASA's NTRS) DTIC Public STINET (DTIC) Data.gov ASDL (initially sponsored by NSF NSDL) NASA: Marshall Technical Reports Server Library of Congress: American Memory NASA Technical Reports Server NASA Gps Environmental & Earth Sciences Information System: GENESIS EPrints Repository@NOAA NSF ATE Central NASA Dryden Technical Reports Server National Agriculture Library Digital Repository DoEd Education Resources Information Center (ERIC) SAO/NASA Astrophysics Data System NASA JPL Beacon eSpace

	National Science Digital Library (NSDL)
Ingest/Harvest:	<i>User Submission; OAI-PMH; OpenDOAR Importer; Celestial</i>

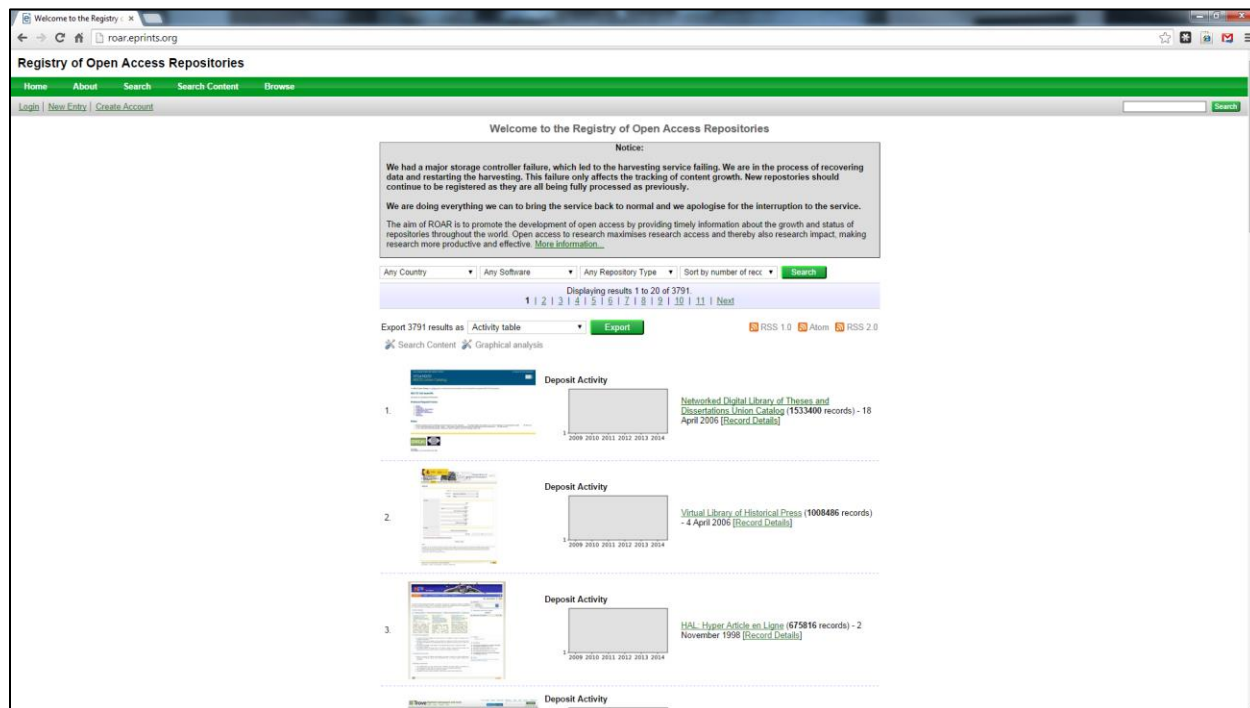


Figure 1: ROAR Homepage

OpenDOAR (The Directory of Open Access Repositories) http://www.opendoar.org/index.html	
<i>An authoritative directory of academic open access repositories. Each OpenDOAR repository has been visited by project staff to check the information that is recorded here. This in-depth approach does not rely on automated analysis and gives a quality-controlled list of repositories.</i>	
Hosted/Maintained by:	SHERPA Services, Centre for Research Communication at the University of Nottingham, UK
Funding by:	JISC; Contributions from University of Nottingham
Part of :	SHERPA Services
Level of Content:	Points to Repository Web site
Listed Metadata: *Filter options	Repository Name Repository URL Description Repository Type* Organisation Org. URL Town/City Country* Location – Latitude. Longitude Subject* Software Platform* Size Content Languages* Policies Remarks OAI Base URL OpenDOAR ID
Federal Repositories Included:	Library of Congress – American Memory Library of Congress – National Jukebo Institute of Museum and Library Services NASA Dryden Technical Reports Server NASA Johnson Technical Reports Server NASA JPL BEACON eSPACE NASA JSC Digital Image Collection NASA JSC Reduced Gravity Program Photographs NASA Marshall Technical Reports Server NASA Technical Reports Server (NTRS) NASA Goddard Library Repository National Agricultural Library Digital Repository (NALDR) NIH PubMed Central NLM Digital Collections NLM PubChem NSF CAUSEweb.org NSF Exploratorium Digital Library National Science Digital Library (NSDL) DoEd Resources Information Center (ERIC) DOE Office of Scientific & Technical Information (OSTI)

Ingest/Harvest:	User Suggestions; OAI-PMH
-----------------	---------------------------

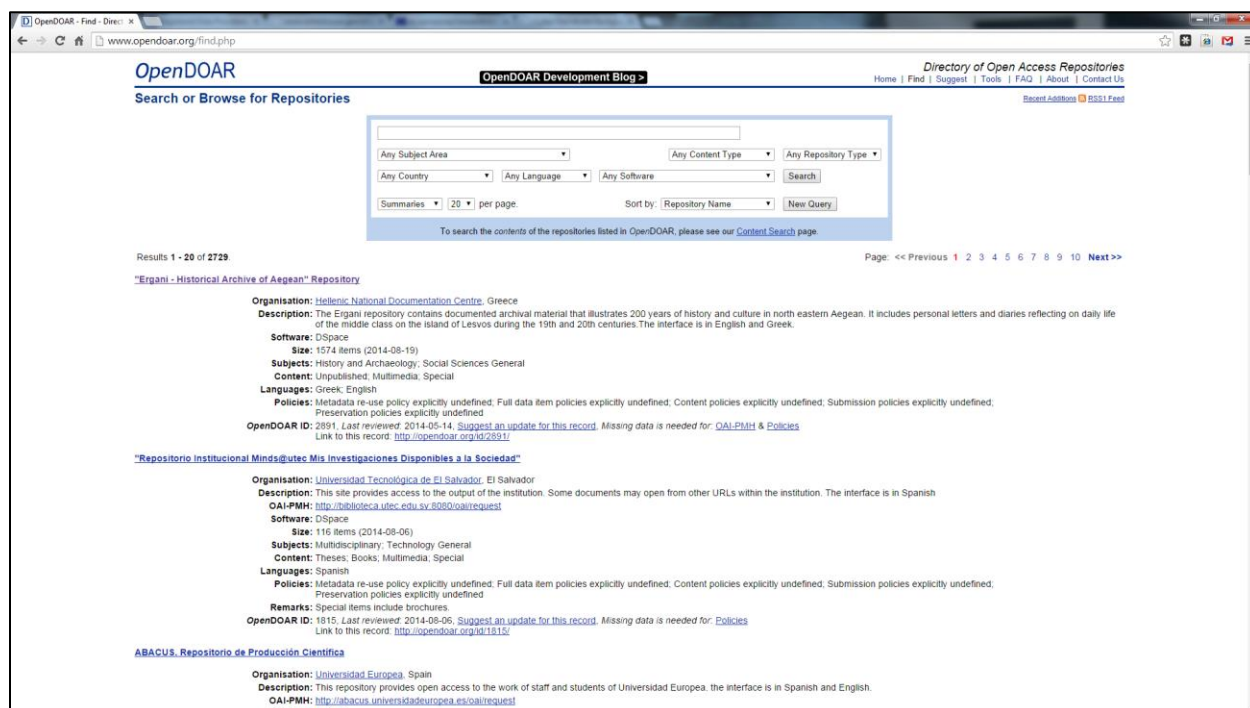


Figure 2: Search or Browse for Repositories in OpenDOAR

Registry of Data Repositories

re3data.org Registry of Research Data Repositories http://www.re3data.org/	
<i>A global registry of research data repositories from different academic disciplines.</i>	
Hosted/Maintained by:	Berlin School of Library and Information Science GFZ German Research Centre for Geosciences Karlsruhe Institute of Technology (KIT) Library
Funding by:	German Research Foundation DFG
Part of :	German Initiative for Network Information (DINI)
Level of Content:	Points to Repository Web site
Listed Metadata: <i>*Filter Options (other options include certificates, open access, persistent identifiers, re3data.org reviewed repositories)</i> (Schema Available)	Name of Repository Additional Name Repository URL Identifier Start Date End Date Subject* Description

	Content type* Country* Keyword Repository type Software Data License Language Size Version Provider type Institution Information Policy Information Database Access Database License Standards Information
Federal Repositories Included:	Global Change Master Directory (GCMD) NASA Distributed Active Archive Center at National Snow & Ice Data Center (NASA DAAC at NSIDC) NASA Atmospheric Science Data Center Clouds and the Earth's Radiant Energy System (CERES) Crystal Dynamics Data Information System (CDDIS) Data.gov DOE Data Explorer NASA Goddard Earth Sciences: Data and Information Services Center (GES DISC) NOAA National Geophysical Data Center National Science Digital Library (NSDL) NOAA Coral Reef Information System (CoRIS) NASA Earth Observing System Data and Information System (EOSDIS) Institute of Museum and Library Services Data Collection (IMLS) NOAA National Climatic Data Center (NCDC) **plus additional US government research data repositories
Ingest/Harvest:	<i>User Suggestions</i>

Suggest a Data Repository

*** Required**

Repository Name *
The full name of the research data repository.

Repository URL *
The URL, which gives reference to the research data repository.

Description *
A textual description containing additional information about the data repository (primary language is English).

Language
The user interface language of the research data repository.
☐ ENG
☐ DE
☐ other

Type
The type of the research data repository.
☐ disciplinary
☐ institutional
☐ other

Size
The number of items contained in the research data repository. Example: 50 datasets; 30 studies (The format is open.)

Start Date
Releasing date of the research data repository.
mm/dd/yyyy

End Date
Date when the research data repository stopped the ingest of new research data (still making the research data available).
mm/dd/yyyy

Suggester's Contact *
Email address of the suggester.

Figure 3: re3data.org Suggest a Data Repository form

Registry of Enterprise Repositories

Digitization Projects Registry (aka Registry of U.S. Government Publication Digitization Projects)

<http://registry.fdlp.gov>

A listing of publicly accessible collections of digitized U.S. Government publications. It was designed to serve as a directory and locator tool of digitization projects as well as to increase awareness and encourage cooperative efforts. At the release of this report, the Digitization Projects Registry is unavailable and is under review by a team at GPO's Library Services and Content Management Department. Their goal is to get updates for many of the records in the registry and remove those that no longer meet the criteria. The ongoing goal is to keep a valuable conversation going while improving and enhancing the Registry.

Hosted/Maintained by:	U.S. Government Printing Office
Funding by:	U.S. Federal Government
Part of :	Federal Depository Library Program
Level of Content:	Points to Digitized Collection Web site
Listed Metadata:	Collection Name Description Contact Information Technical Information Geographic Coverage SuDocs Classification State/Location
Federal Repositories Included:	<i>Website is currently unavailable due to internal security review by U.S.</i>

	<i>Government Printing Office.</i>
Ingest/Harvest:	<i>User submission only</i>

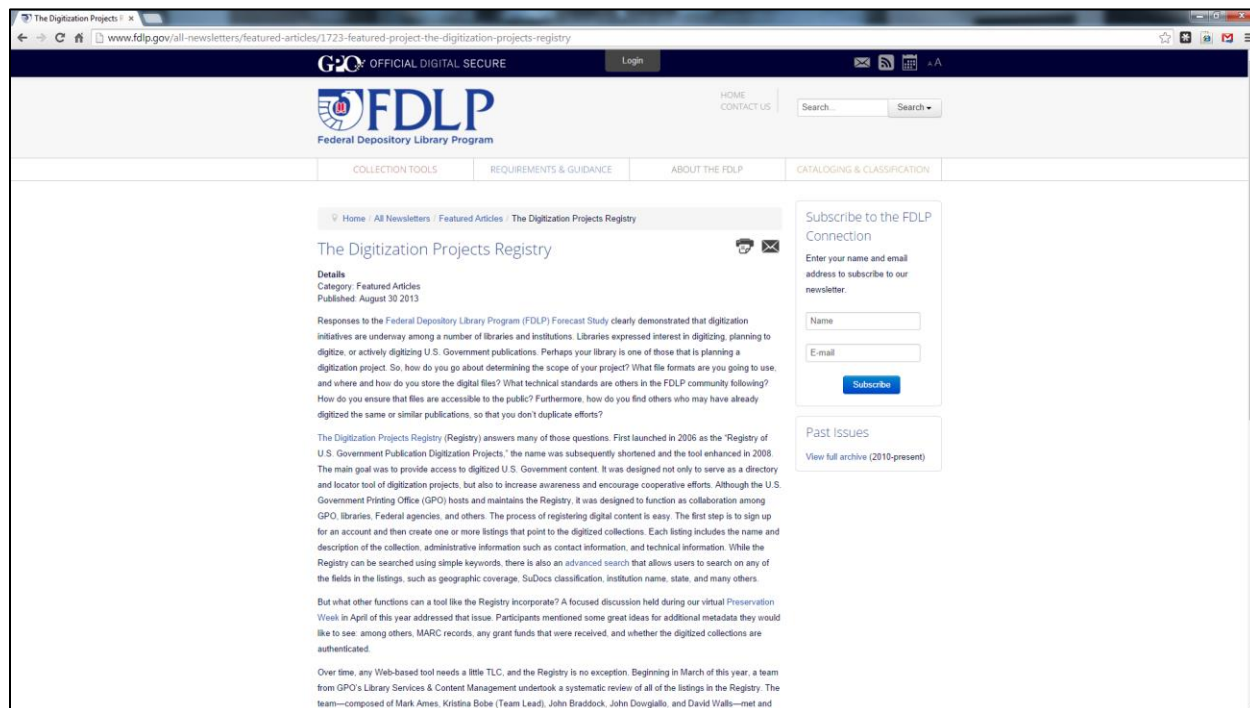


Figure 4: Article on the Digitization Projects Registry

Registries of Repositories Based on a Specific Technology

DuraSpace Registry http://www.duraspace.org/registry	
<i>A list of repository instances that use DSpace, DuraCloud, or Fedora repository technologies.</i>	
Hosted/Maintained by:	DuraSpace
Funding by:	DuraSpace Organization
Part of :	DuraSpace
Level of Content:	Points to Repository Web site
Listed Metadata:	Repository Name Repository Technology Country Type of Institution Repository access Version Operating System/Platform Relational Database Use cases Type of Content Integrations/customizations

	ID
Federal Repositories Included:	BEACON eSpace (NASA JPL) National Agriculture Library Digital Collections Government Funded Technical Reports Repository (NTIS) Goddard Library Repository (NASA) NASA Langley Research Center NLM Digital Collections U.S. CDC Web site U.S. Geological Survey Web site U.S. National Library of Medicine
Ingest/Harvest:	<i>User submission; DuraSpace entry due to use of Fedora, DSpace, or DuraCloud</i>

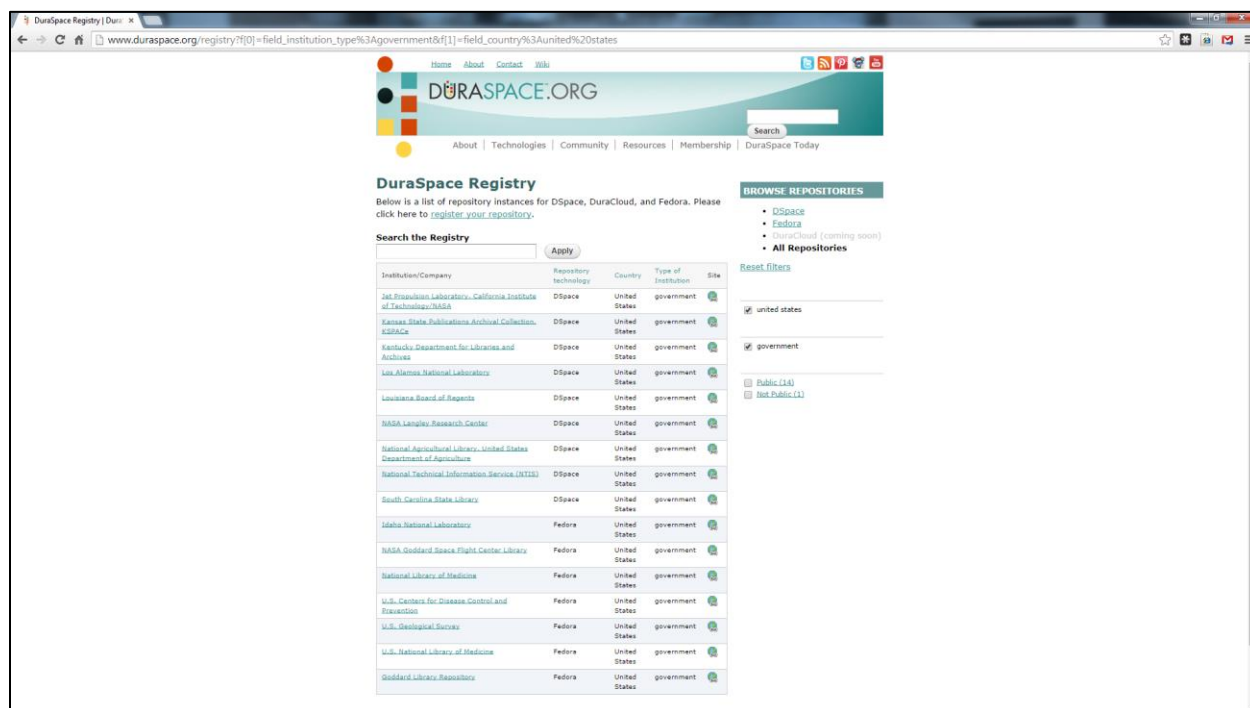
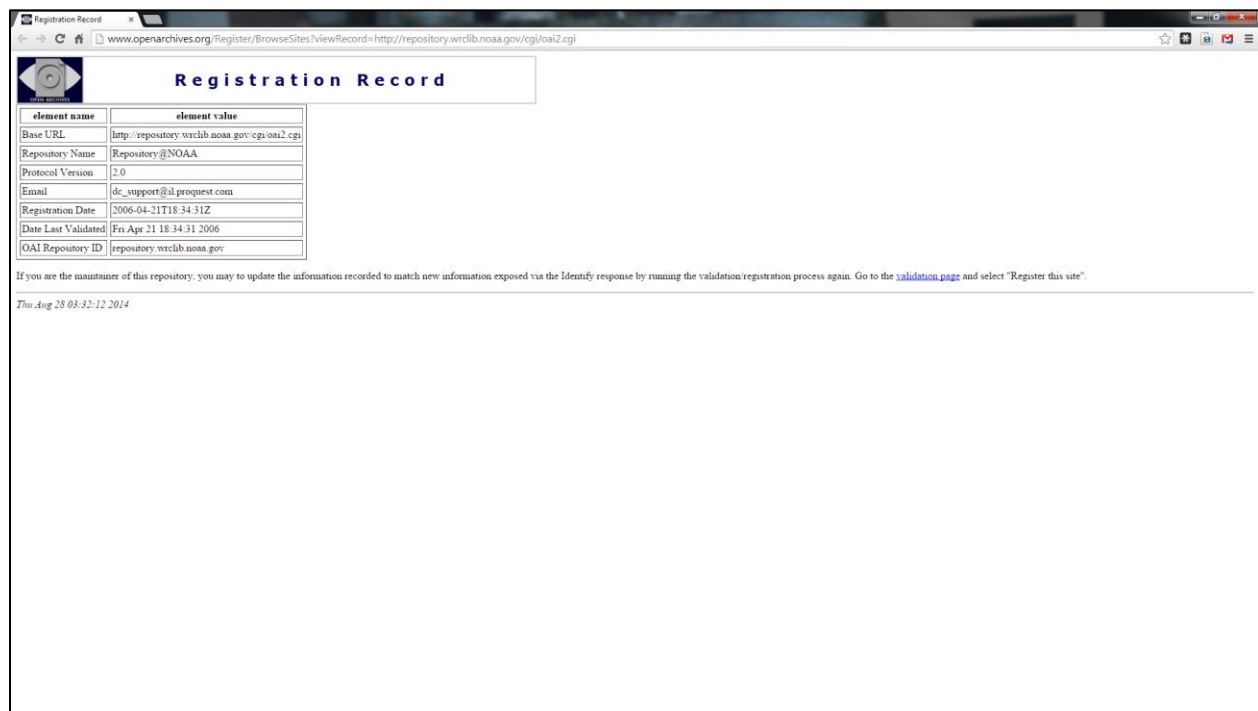


Figure 5: DuraSpace Registry - Filtered to show U.S. government repositories

OAI Registered Data Providers http://www.openarchives.org/Register/BrowseSites	
<i>A list of registered Open Archives Initiative (OAI) conforming repositories. The data providers (repositories) are registered through the OAI registration and validation page.</i>	
Hosted/Maintained by:	Cornell University Library Information Technology
Funding by:	Andrew W. Mellon Foundation Coalition for Networked Information Digital Library Federation Microsoft Corporation Alfred P. Sloan Foundation National Science Foundation

Part of :	Open Archives Initiative (OAI)
Level of Content:	Points to Repository Web site and OAI-identifier namespace
Listed Metadata:	Base URL Repository Name Protocol Version Email Registration Date Data Last Validated OAI Repository ID
Federal Repositories Included:	EPA OAI Archive Library of Congress Open Archive Initiative Repository 1 Langley Technical Reports Server (NASA) NACA Profiles in Science (NLM) PubMed Central (NLM) Repository@NOAA
Ingest/Harvest:	<i>OAI entry and update through the OAI registration and validation process</i>



element name	element value
Base URL	http://repository.wrclib.noaa.gov/cgi/oai2.cgi
Repository Name	Repository@NOAA
Protocol Version	2.0
Email	dc_support@il.proquest.com
Registration Date	2006-04-21T18:34:31Z
Date Last Validated	Fri Apr 21 18:34:31 2006
OAI Repository ID	repository.wrclib.noaa.gov

If you are the maintainer of this repository, you may update the information recorded to match new information exposed via the Identify response by running the validation registration process again. Go to the [validation page](#) and select "Register this site".

Thu Aug 28 03:32:12 2014

Figure 6: OAI Data Provider – Repository Record

Appendix B: Metadata by Registry

Listed Metadata*	ROAR	OpenDOAR	re3data.org	DPR	Duraspace	OAI
Title						
Repository Name	X	X	X		X	X
Additional Name			X			
Collection Name				X		
Type						
Repository Type	X	X	X			
Content Type		X	X		X	
Identifier			X			
Repository URL		X	X	X	X	
Homepage	X					
Base URL						X
ID					X	
ROAR ID	X					
OAI Repository ID						X
OpenDOAR ID		X				
Coverage						
City/State		X		X		
Country	X	X	X		X	
Location	X	X		X		
Geographic Coverage				X		
Organization	X	X				
Org./Institution Name			X			
Additional Name			X			
Org. URL		X	X			
Contact			X	X		
Email						X
Institution Type			X		X	
Description		X	X	X		
Subject	X	X	X			
Keyword			X			
Classification				X		
Software	X		X			
Relational Database					X	
Software Platform		X				
Operating System/ Platform					X	
Repository Technology					X	
Language		X	X			
Date						
Birth Date	X					
Start Date			X			

Registration Date						X
End Date			X			
Date Last Validated						X
Size		X	X			
Record Count	X					
Version			X		X	
Protocol Version						X
Access						
Repository Access					X	
Database Access			X			
License						
Database License			X			
Data License			X			
Policy Information		X	X			
Record Creator	X					
Compatibility						
Standards			X			
Integration/customization					X	
Harvester						
OAI-PMH	X					
OAI Base URL		X				X
Data and/or Service provider			X			
Other Registry Link	X					
Notes						
Use Cases					X	
Remarks		X				

*There could be additional metadata fields available for each registry that are only available to account users. These fields were publically accessible on the site.